Exhibit 2

the benzyloxycarbonyl group. Digestion of a 30-residue peptide beginning Gly-Pro-HyPro- with this new enzyme yielded Gly-Pro only, an indication of its specificity and ability to act on longer polypeptides. Presumably, a similar enzyme for use in the DCP system, when necessary, should also become available at some future date.

In summary, the dipeptidyl peptidase approach to polypeptide sequence analysis is a useful alternative to other methodologies for these determinations. With assets of speed and sensitivity, as well as the ability to proceed from either the NH_2 or COOH terminus, it should be the method of choice in a number of situations.

[47] Establishing Homologies in Protein Sequences

By MARGARET O. DAYHOFF, WINONA C. BARKER, and LOIS T. HUNT

That different species contain homologous proteins was known long before the exact chemical sequences had been elucidated. Although it was clear that the homologous structures were not identical, nevertheless, mixed systems that functioned perfectly well chemically could be constructed with enzymes from different species. The results from protein sequence determinations over the last 30 years have made clear the nature of the homologous structures. There is an ongoing process of mutation and selection whereby a normal protein of a species can change from one form to a slightly altered form. The accepted mutations are of two principal kinds: point mutations, including alteration of one nucleotide of the triplet coding for one amino acid and deletions or insertions of one or a few whole codons; and large changes in the amount of genetic material, believed to be caused by unequal crossing-over of the chromosomes, resulting in duplications or deletions that can include entire genes. When gene pools become isolated, through either a separation of interbreeding populations or a duplication of genetic material within a species, the copies gradually acquire changes independently of one another. At first the sequences are so similar that there is no question about their common origin. With increasing time more and more change occurs until it may no longer be possible to recognize the similarity.

Frequently the biochemist is confronted with the problem of identifying a newly determined protein sequence or a protein sequence inferred from a nucleotide sequence. If proteins are less than 30% different from each other, then similarity can often be detected immunologically. DNA coding regions can be identified by annealing if the nucleotide sequences

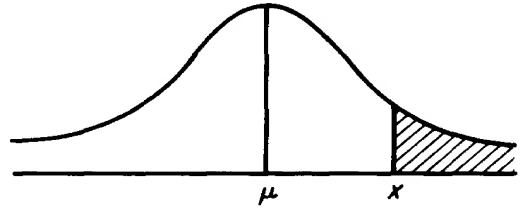
ISBN 0-12-181991-4

are less than 30% different. The identification of relationships between proteins that are up to 75 or 80% different can be accomplished by comparison of the sequences.

In this chapter we will be particularly concerned with statistical tests capable of illuminating even very distant relationships. These tests are based on the hypothesis that the sequence under consideration and another selected sequence are more similar than one would expect by chance. Sequences can be selected for consideration by many criteria. Frequently they are chosen because they are similar in some aspects, for example, chemical function, active site, prosthetic group, unusual modification of a residue, tertiary structure, interaction with other molecules, amino acid composition, immunological similarity, physiological activity, or because of strong similarity of short sequence segments. They can also be selected by examining sequences for known active sites, by comparing parts of sequences to tabulations of sequences or to an alphabetized listing of the sequences of known segments, or by performing a computer search of a segment against all the known sequences.

Two types of statistical tests are in common use. In one, all the segments of a given length in one sequence are compared with all the segments in the second sequence.[1-3] In the other, the best alignment of two sequences is made.[1,4-8] We have based our computer programs RELATE and ALIGN on these two methods.[1,4] In both methods, a scoring matrix is required and a numerical property of the comparison is calculated. This same property is also calculated for a large number of pairs of permuted sequences (with the same compositions as the real sequences). The mean and standard deviation of the property are estimated from the distribution of scores of the permuted sequences. An assessment of the probability of the real score occurring by chance can then be made on the basis of the probabilities of standardized scores for the normal distribution (Fig. 1). These methods focus on the pattern in the sequence and do not include any contribution from similarity in the amino acid composition of the proteins. In sequences of nearly average composition, the contribution of

[1] M. O. Dayhoff, *in* "Atlas of Protein Sequence and Structure" (M. O. Dayhoff, ed.), Vol. 5, Suppl. 3, pp. 1–8. National Biomedical Research Foundation, Washington, D.C., 1979.
[2] W. M. Fitch, *J. Mol. Biol.* **16,** 9 (1966).
[3] W. M. Fitch, *J. Mol. Biol.* **49,** 1 (1970).
[4] W. C. Barker and M. O. Dayhoff, *in* "Atlas of Protein Sequence and Structure 1972" (M. O. Dayhoff, ed.), Vol. 5, pp. 101–110. National Biomedical Research Foundation, Washington, D.C., 1972.
[5] S. B. Needleman and C. D. Wunsch, *J. Mol. Biol.* **48,** 443 (1970).
[6] P. H. Sellers, *SIAM J. Appl. Math.* **26,** 787 (1974).
[7] P. H. Sellers, *Proc. Natl. Acad. Sci. U.S.A.* **76,** 3041 (1979).
[8] T. F. Smith, M. S. Waterman, and W. M. Fitch, *J. Mol. Evol.* **18,** 38 (1981).



Normal distribution with mean μ and standard deviation σ

The probability of obtaining a score greater than or equal to x is shown in terms of z, the number of standard deviation units from x to the mean, $z = (x - \mu)/\sigma$.

Probability of a Score $\geq$ x	z (SD units)	z (SD units)	Probability of a Score $\geq$ x
10^{-1}	1.28	0.0	0.500
10^{-2}	2.33	0.5	0.309
10^{-3}	3.09	1.0	0.159
10^{-4}	3.72	1.5	0.668×10^{-1}
10^{-5}	4.26	2.0	0.227×10^{-1}
10^{-6}	4.75	2.5	0.621×10^{-2}
10^{-7}	5.20	3.0	0.135×10^{-2}
10^{-8}	5.61	3.5	0.233×10^{-3}
10^{-9}	6.00	4.0	0.317×10^{-4}
10^{-10}	6.36	4.5	0.340×10^{-5}
10^{-15}	7.94	5.0	0.287×10^{-6}
10^{-20}	9.26	5.5	0.190×10^{-7}
		6.0	0.987×10^{-9}
		6.5	0.402×10^{-10}
		7.0	0.128×10^{-11}
		8.0	0.622×10^{-15}
		9.0	0.113×10^{-18}
		10.0	0.762×10^{-23}

FIG. 1. Probabilities of standardized scores for the normal distribution. This figure was taken, with permission, from Table 35 in the "Atlas of Protein Sequence and Structure," Vol. 5, Suppl. 3, p. 374, 1979.

composition is weak. For sequences of uncommon composition, the correct null hypothesis is not well understood. For example, if polylysine were found in two organisms, would this represent common ancestry or would it derive from poly(A) being incorporated into genes in two entirely separate events? Methods for inferring relationships from amino acid compositions have been studied[9–14]; we will not consider them here. In

[9] J. J. Marchalonis and J. K. Weltman, *Comp. Biochem. Physiol. B* **38B**, 609 (1971).
[10] C. E. Harris and D. C. Teller, *J. Theor. Biol.* **38**, 347 (1973).
[11] A. Cornish-Bowden and A. Marson, *J. Mol. Evol.* **10**, 231 (1977).
[12] A. Cornish-Bowden, *Biochem. J.* **191**, 349 (1980).
[13] H. M. Shapiro, *Biochim. Biophys. Acta* **236**, 725 (1971).
[14] H. Vogel, *J. Mol. Evol.* **6**, 271 (1975).

any case, the overall statistical probability would be derived by multiplying that from composition information and that derived from sequence pattern, since these properties are independent.

Selection Using Printed Tabulations

In scanning known sequences by eye, one usually concentrates on regions containing residues that are most highly conserved: Cys, Trp, and the two residues Tyr and Phe (which substitute principally for one another). Where these residues match, the rest of the residues should be examined. If the relationship is so strong that 40% of the residues are identical over 25 residues or more, without introducing breaks or gaps in either sequence, the relationship is quite definite and statistical tests are not necessary. Sequences more than 50% different usually require one or more breaks to align them for maximum homology. One usually must resort to statistical tests to evaluate the more distant relationships; unrelated sequences are only 72 ± 6% different, excluding positions with a gap, when an unlimited number of gaps are permitted and the penalty for making a break in either sequence is equal to the score for one match.

The "Protein Segment Dictionary," a key-amino-acid-in-context listing of all 15-residue segments from the sequences known in 1977, alphabetized on the sixth and following amino acids, can be referenced for exact matches to short sequences of three or more amino acids (Fig. 2).[15] Because fewer than 2% of all possible sequences of five amino acids and 0.2% of those sequences six long actually occur in this collection of se-

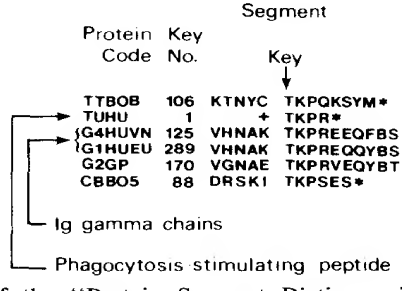


FIG. 2. A portion of the "Protein Segment Dictionary."[15] The sequence of the phagocytosis-stimulating peptide is completely contained in the sequences of the immunoglobulin γ chains and nowhere else in the collection. This suggests that the peptide may be derived chemically from the γ chain, particularly since it is known to function in association with the γ chain.

[15] M. O. Dayhoff, L. T. Hunt, W. C. Barker, R. M. Schwartz, and B. C. Orcutt, "Protein Segment Dictionary 78." National Biomedical Research Foundation, Washington, D.C., 1978.

quenced proteins, a search for a specific penta- or hexapeptide will usually turn up the source protein and its close relatives or nothing at all.

The new sequence should be examined for known active sites. These have often been conserved over large phylogenetic distances. For example, the sequence Gly-Asp-Ser-Gly-Gly surrounding the serine active site of trypsin is absolutely conserved in all 19 of the sequenced serine proteases related to trypsin, including four bacterial sequences, and it occurs nowhere else in our present protein sequence database.

Identification of Very Similar Sequences

In identifying short segments that are identical with or very similar to a test piece, as in the problem of identifying the source of a peptide believed to originate by chemical degradation of a larger protein, a search using the unitary (or identity) matrix is appropriate. Alternatively, the sequence can be looked for in the "Protein Segment Dictionary."[15]

For a segment identification to be possible, the residues that have been sequenced need not be contiguous and the amino acids do not need to be unambiguously identified. However, such searches are usually practical only with a computer. An exact match of from six to nine amino acid residues (depending on the frequency of occurrence of these amino acids) should suffice to identify uniquely a segment among all human sequences.[16]

Computer Methods: The Database

When relationships are not immediately obvious one can use computer methods to compare the new sequence to the other known sequences, which are conveniently accessible online in the Protein Sequence Database.[17] In this database, we have organized the known sequences into hierarchical groups of superfamilies, families, subfamilies, and entries.[18] The number in each group, the criteria for clustering, and the method of identification of the hierarchical levels is shown in Table I. The sequences are clustered into superfamilies of sequences that can be shown to be significantly related by statistical methods. Within the superfamily,

[16] M. O. Dayhoff and B. C. Orcutt, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 2170 (1979).

[17] Protein Sequence Database (M. O. Dayhoff, L. T. Hunt, W. C. Barker, B. C. Orcutt, L.-S. Yeh, H. R. Chen, D. G. George, M. C. Blomquist, J. A. Fredrickson, and G. C. Johnson). National Biomedical Research Foundation, Washington, D.C., 1981.

[18] M. O. Dayhoff, W. C. Barker, L. T. Hunt, and R. M. Schwartz, *in* "Atlas of Protein Sequence and Structure" (M. O. Dayhoff, ed.), Vol. 5, Suppl. 3, pp. 9–24. National Biomedical Research Foundation, Washington, D.C., 1979.

TABLE I
SUPERFAMILY ORGANIZATION

Number in group	Group	Criteria for clustering	Number of residues in first sequences of each group
512	Superfamilies	Probability of similarity by chance $<10^{-6}$	98,394
774	Families	<50% different	142,045
1161	Subfamilies	<20% different	192,097
1667	Entries	<5% different	258,126

we have divided the sequences into families. Proteins within a family usually differ at fewer than half of their amino acid positions; their similarity of function has often been recognized before the sequences were known, and they have identical or very similar names. The sequences have been further divided into subfamilies. Sequences within a subfamily usually differ from each other at fewer than 20% of their amino acid positions. Within a subfamily, sequences that differ by less than 5% are usually described in a single database entry. If the rate of change in amino acid sequences were exactly proportional to time, such a clustering would identify the twigs, branches, and boughs of the evolutionary tree. The rate has been only approximately constant, and so, in a clustering procedure such as this, there will always be cases that are borderline, some pairs within a group being below the cutoff and some above. In spite of difficulties in a few details, Table I gives a very good impression of the sequence information that is available. At present the number of newly investigated superfamilies is doubling in about 3 years and the number of newly sequenced residues is doubling in about 4 years.

The clustering procedure also gives a basis for selecting sequences for a given purpose. For a minimal database for computer searches, it would be adequate to select one sequence from each family and thereby reduce the costs of searching by a factor of 2.

We postulate that there may be only 1000 superfamilies of functional proteins, some containing several groups of very distantly related proteins. The probability, 10^{-6}, for clustering sequences into a superfamily has been chosen so that, in making all possible comparisons of 2000 sequences, each representing one group of clearly related sequences (4×10^6 comparisons), the number of groups placed into the wrong superfamily will be very small, approximately 4. If one used a probability of 10^{-5} for significance, one would expect about 40 misclassifications, or an error in 4% of the groups. As more information becomes known on the charac-

teristics of the superfamilies, such misclassified sequences will be evident. Also, as more sequences are known for any one superfamily, the pattern of conserved residues will become clearer. Incorporating this information into the statistical tests, through weighting of the conserved residues, will increase the detectability of related sequences.

Our superfamily definition is based on sequence information so that a relationship can be readily verified by objective methods. Some of the superfamilies may have had a very distant common evolutionary origin, which may be demonstrated on the basis of other evidence, such as the positions of α-carbon atoms from X-ray crystallography.[19,20] Many of the problems involved in the statistical detection of very distantly related sequences have recently been discussed by Doolittle.[21]

Searching the Database

In investigating a new sequence of unknown relationship to form hypotheses of evolutionary similarity to other sequences in the database, we first use the computer program SEARCH.[1] We select several test segments of, for example, 25-residue length from different regions of the new sequence. Each of these is compared with all 25-residue segments and the shorter end segments of each known sequence. A score for the segments is accumulated by adding the pair scores of each amino acid in the segment searched with the corresponding amino acid in each segment of the database. The pair scores are contained in a matrix. Several scoring systems, summarized in matrices, are in common use (see below). The simplest system assigns a score of 1 for identities and 0 for nonidentities. Table II shows the average numbers of identical residues found in unrelated segments in a search of the database. Segments were 25 residues in length, and 218,000 segments were compared in each of 15 searches. From these numbers we derived the number of high-scoring segment pairs to be expected in a RELATE run and the probability of finding a pair of segments with at least a given number of identities. For sequences that are related, the similarity extends beyond the segment searched, whereas for unrelated sequences it usually does not.

For detecting distant relationships, we have found the mutation data matrix to be best[22] (see Fig. 3). The distribution of scores from unrelated

[19] P. Keim, R. L. Heinrikson, and W. M. Fitch, *J. Mol. Biol.* **151,** 179 (1981).
[20] S. J. Remington and B. W. Matthews, *Proc. Natl. Acad. Sci. U.S.A.* **75,** 2180 (1978).
[21] R. F. Doolittle, *Science* **214,** 149 (1981).
[22] R. M. Schwartz and M. O. Dayhoff, *in* "Atlas of Protein Sequence and Structure" (M. O. Dayhoff, ed.), Vol. 5, Suppl. 3, pp. 353–358. National Biomedical Research Foundation, Washington, D.C., 1979.

TABLE II
IDENTITIES IN UNRELATED 25-RESIDUE SEGMENTS (NO GAPS PERMITTED)

Number of identities per segment	Average number of segments in a database search	Number of pairs expected in 10^4 RELATE scores	Probability for one pair
≧7	87	4	4.0×10^{-4}
≧8	11	0.5	4.9×10^{-5}
≧9	1	0.05	4.9×10^{-6}
≧10	0.02	0.001	8.4×10^{-8}

C	Cys	12																			
S	Ser	0	2																		
T	Thr	-2	1	3																	
P	Pro	-3	1	0	6																
A	Ala	-2	1	1	1	2															
G	Gly	-3	1	0	-1	1	5														
N	Asn	-4	1	0	-1	0	0	2													
D	Asp	-5	0	0	-1	0	1	2	4												
E	Glu	-5	0	0	-1	0	0	1	3	4											
Q	Gln	-5	-1	-1	0	0	-1	1	2	2	4										
H	His	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R	Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K	Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M	Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I	Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L	Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V	Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F	Phe	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y	Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W	Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
		Cys	Ser	Thr	Pro	Ala	Gly	Asn	Asp	Glu	Gln	His	Arg	Lys	Met	Ile	Leu	Val	Phe	Tyr	Trp

FIG. 3. The mutation data matrix, or log odds matrix for 250 PAMs (see text). Elements are shown multiplied by 10. The neutral score is 0. A score of −10 means that the pair would be expected to occur only one-tenth as frequently in related sequences as chance would predict, and a score of +2 means that the pair would be expected to occur 1.6 times as frequently. The order of the amino acids has been arranged to illustrate the patterns in the mutation data. This figure was taken, with permission, from Fig. 84 in the "Atlas of Protein Sequence and Structure," Vol. 5, Suppl. 3, p. 352, 1979.

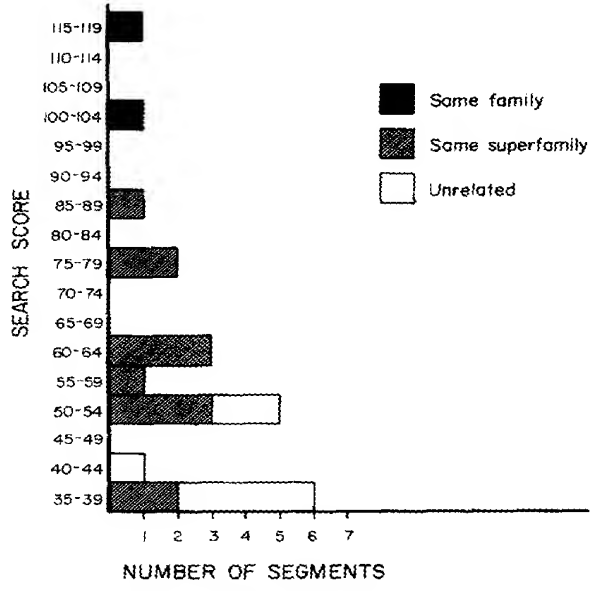


FIG. 4. Histogram of scores for a search of bovine trypsinogen. The 25-residue segment used, positions 34 to 58, contains the histidine active site. Two distantly related eukaryote sequences, factor X and haptoglobin β chain, did not get high scores in this search.

segments is approximately normal; related segments appear in an abnormally long tail of high scores. Typically, for a 25-residue segment, all corresponding sequences of the same family (<50% different) appear above the distribution of scores of unrelated segments (see Fig. 4). About half of the more distantly related sequences in the same superfamily are also above this distribution, whereas the rest may be within the upper tail of scores from unrelated segments. If the initial searches produce no scores above 50, probably there are no other sequences from the same family in the data collection. Searches using a very short segment, say 10 residues, produce a distribution of unrelated scores having a very long tail of high scores that often obscures the real relationships. For longer segments, say 40 residues, there are usually insertions or deletions in the sequences that interfere with obtaining high scores.

In cases where no obvious relationship is found in a computer search, one may compare the top-scoring 30–40 segments with the test segment using several other criteria. For example: Are identical residues clustered (4–6 together)? Would the introduction of a single gap greatly improve the match? Does a matching segment occur in the same part of the molecule as the test segment, or is there the possibility of internal duplications if the

regions are different? Do the test protein and the matching protein have similar secondary or tertiary structures or similar functions or functional domains? Do test segment and matching segment share identities that bind the same functional moieties? Are both from the same organelle, cell type, or major group of organisms? Positive answers to such questions can increase the possibility of a relationship and indicate candidates for ALIGN or RELATE analyses.

If the initial examination does not identify a relationship, we may search more exhaustively. Typically we search each successive segment of 25 residues. This provides for detection even in the case where only a fragment of a related sequence is present in the database. We will consider further two kinds of screening procedures. In the first, especially suitable for evaluation of search results by eye, all sequences that get a high score, ≥ 40, on at least one search are selected for statistical evaluation with programs ALIGN and RELATE. In the second, sequences containing a moderately high score, ≥ 30, on two or more searches are selected.

In any screening process, one must balance the probability and value of success against the cost of examining spurious suggestions. We conjecture that, at present, between 25 and 50% of all of the superfamilies to be found in organisms are represented in the database. If so, in more than half of the cases no related sequence will be found even with perfect detection methods. Even where there are very distant relationships that could be confirmed by statistical tests on the complete sequences, it may be very expensive or even impossible to find them by searching for short segments. Success depends on the related sequence containing enough regions giving high scores that these will be observed in the searches performed.

For this chapter we searched a total of 101 segments from 13 sequences: bovine trypsinogen,[23] *Chromatium vinosum* high-potential iron–sulfur protein,[24,25] *Escherichia coli* thioredoxin,[26] *Streptomyces erythreus* ribonuclease,[27] horse alcohol dehydrogenase,[28] *Desulfovibrio vulgaris* flavodoxin,[29] *E. coli* K12 dihydrofolate reductase type I,[30] human antithrombin-III,[31] human α_1-microglobulin,[32] *E. coli* 50S ribosomal protein L7/L12,[33,34] bovine α-crystallin A chain,[35] human prolactin,[36] and human thyrotropin α chain.[37,38] For each search, the highest score re-

[23] O. Mikes, V. Holeysovsky, V. Tomasek, and F. Sorm, *Biochem. Biophys. Res. Commun.* **24,** 346 (1966).

[24] S. M. Tedro, T. E. Meyer, R. G. Bartsch, and M. D. Kamen, *J. Biol. Chem.* **256,** 731 (1981).

[25] K. Dus, S. Tedro, and R. G. Bartsch, *J. Biol. Chem.* **248,** 7318 (1973).

[26] A. Holmgren, *Eur. J. Biochem.* **6,** 475 (1968).

[27] N. Yoshida, A. Sasaki, M. A. Rashid, and H. Otsuka, *FEBS Lett.* **64,** 122 (1976).

[28] H. Jornvall, *Eur. J. Biochem.* **16,** 41 (1970).

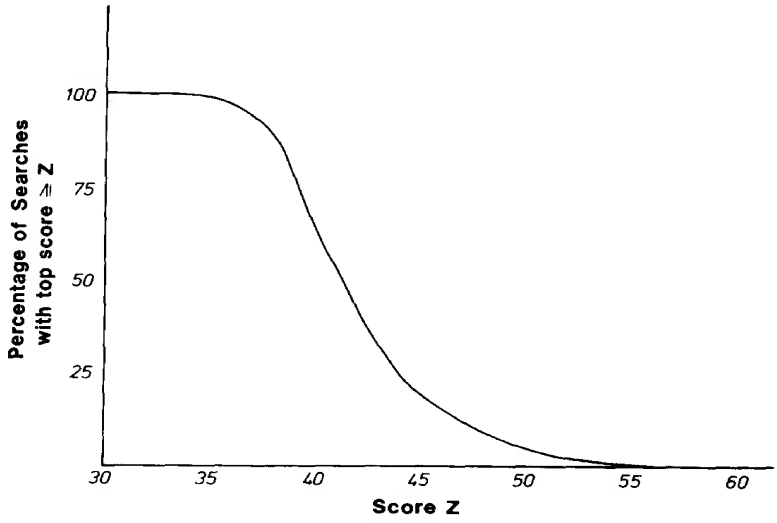


FIG. 5. Percentage of searches with a top score $\geq Z$. This curve is for a database of 250,000 residues. As the database increases in size, the curve will shift to the right. One hundred and one segments of 25 residues, drawn from 13 sequences, were searched using the mutation data matrix.

ceived by an unrelated segment was tallied. Figure 5 shows the percentage of searches containing a score as high or higher than a given value. The distribution of highest scores is approximately normal, with a mean of 41.5 and a standard deviation of 4.1. On this basis, the probability of the top score being above 60 is $<3.4 \times 10^{-6}$ and above 70 is 1.3×10^{-11} in

[29] M. Dubourdieu and J. L. Fox, *J. Biol. Chem.* **252,** 1453 (1974).
[30] D. R. Smith and J. M. Calvo, *Nucl. Acids Res.* **8,** 2255 (1980).
[31] T. E. Petersen, G. Dudek-Wojciechowska, L. Sottrup-Jensen, and S. Magnusson, *in* "The Physiological Inhibitors of Blood Coagulation and Fibrinolysis" (D. Collen, B. Wiman, and M. Verstraete, eds.), pp. 43–54. Elsevier/North-Holland Biomedical Press, Amsterdam, 1979.
[32] T. Takagi, K. Takagi, and T. Kawai, *Biochem. Biophys. Res. Commun.* **98,** 997 (1981).
[33] C. Terhorst, W. Moeller, R. Laursen, and B. Wittmann-Liebold, *Eur. J. Biochem.* **34,** 138 (1973).
[34] L. E. Post, G. D. Strycharz, M. Nomura, H. Lewis, and P. P. Dennis, *Proc. Natl. Acad. Sci. U.S.A.* **76,** 1697 (1979).
[35] F. J. van der Ouderaa, W. W. de Jong, and H. Bloemendal, *Eur. J. Biochem.* **39,** 207 (1973).
[36] N. E. Cooke, D. Coit, J. Shine, J. D. Baxter, and J. A. Martial, *J. Biol. Chem.* **256,** 4007 (1981).
[37] J. C. Fiddes and H. M. Goodman, *Nature (London)* **281,** 351 (1979).
[38] M. R. Sairam and C. H. Li, *Can. J. Biochem.* **55,** 755 (1977).

searches of the approximately 220,000 unrelated 25-residue segments in the current database. Such high scores are not independent of amino acid composition; they are found more often for segments that contain several Tyr, Phe, Trp, or Cys residues.

In comparing one single segment with another, the probability of obtaining a score above 30 ($P = 1.3 \times 10^{-4}$) or above 40 ($P = 5.9 \times 10^{-6}$) can be estimated from the number of such scores obtained in the searches. There were 513 scores above 30 and 24 scores above 40 obtained from sequences not known to be related, in comparing almost 4 million segments in 18 searches. Typically, for the present database, in searching ten 25-residue segments, about 13 sequences will be identified with scores greater than 40, 68 with scores over 35, and perhaps 285 with scores over 30. Sequences with the highest scores in one or more of the searches can be examined by eye for more extensive similarity. With a simple computer program one can locate sequences with two or more fairly high-scoring segments in approximate register. Very few such sequences are found in examining all scores of 30 or higher.

The possibility of finding a related sequence by getting a high SEARCH score was also examined. In comparing all segments from the sequence searched with all those from the related sequence, the number of scores, S, greater than or equal to a certain value, Z, was counted. The number of segments of the first sequence that could be chosen for a search is the length, L, minus 24. The probability, $P(Z)$, of finding a high score in one search of an arbitrarily chosen segment is then given by $P(Z) = S/(L - 24)$. The decision as to how many searches to make can be reduced to a sampling problem based on this probability.

For the 13 sequences that were searched, there were 28 cases in which a sequence from a different family was present in the database and the relationships could be detected using our computer programs ALIGN or RELATE, described below. The 28 pairs of related sequences were grouped by percentage difference. Table III shows the probabilities of a segment giving scores equal to or greater than 40 and 30 for the related sequences that we examined. These probabilities are calculated from the average percentage of the segments that would give a sufficiently high SEARCH score.

From these probabilities, one can readily calculate the probability of finding at least one segment with a score ≥ 40, or two or more segments with a score ≥ 30, in performing a specified number of searches. These are the probabilities of the success of the screening procedures when a related sequence is available. In Table IV we show the values for finding one or more scores, P_N, and two or more scores, P_N^D, in making $N = 5$ and $N = 10$ searches. If the screen involves selecting sequences with two scores ≥ 30

TABLE III
PROBABILITY OF A HIGH SEARCH SCORE AS A FUNCTION OF SEQUENCE DIFFERENCE

% Difference[a]	$P(30)$[b]	$P(40)$
50–55	—	—
55–60	0.69	0.53
60–65	0.59	0.37
65–70	0.40	0.22
70–75	0.47	0.27
75–80	0.28	0.10

[a] Sequences were aligned using the mutation data matrix with a bias of 6 and a gap penalty of 6. The percentage difference was formed by counting the number of positions with identical residues divided by the total number of positions containing residues in both sequences. Positions with a gap in either sequence were ignored.

[b] $P(30)$ is the probability that a score of 30 or greater will be obtained for the homologous segment on a single search.

TABLE IV
PROBABILITY[a] OF SUCCESS IN THE SEARCH SCREENING PROCEDURE

P	P_5	P_{10}	P_5^D	P_{10}^D
0.1	0.41	0.65	0.08	0.26
0.2	0.67	0.89	0.26	0.63
0.3	0.83	0.97	0.47	0.85
0.4	0.92	0.99	0.66	0.95
0.5	0.97	1.00	0.81	0.99
0.6	0.99	1.00	0.91	1.00
0.7	1.00	1.00	0.97	1.00
0.8	1.00	1.00	0.99	1.00
0.9	1.00	1.00	1.00	1.00
1.0	1.00	1.00	1.00	1.00

[a] P is the probability that the related segment will get a high score on a single search. P_N is the probability that a related segment will get a high score on one or more of N searches. P_N^D is the probability that the related segments will get high scores on two or more of N searches.

in making 10 searches, then the probability of finding a related sequence about 60–65% different [$P(30) = 0.59$] is 1.00. If it involves selecting sequences with one score ≥ 40 [$P(40) = 0.37$], then the probability of finding a related sequence is 0.98. The probability of missing a relationship is $\leq 2\%$. From Table IV it is readily seen that the probability of success of the screening procedure (when related segments are present) depends on the number of segments searched as well as on the probability of success in any one search.

Alignment Scores: Program ALIGN

If the results of the screening procedure using the SEARCH program suggest sequences that might be homologous, we proceed with simple objective methods to find the probability that the similarity occurred by chance. Program ALIGN[1,4] is particularly suitable for sequences of similar length with matching segments positioned comparably. For any given alignment between two sequences a numerical value can be computed. The contribution of each match of a residue in one sequence with one in the other is defined by a matrix of values. A break may be introduced in either sequence, for which a gap penalty is incurred. The score for an amino acid matching a gap position is zero. The score for an alignment is the sum of scores for the positions along the alignment and the gap penalties incurred. Considering all possible alignments of the residues and any number of gaps, the basic algorithm of program ALIGN determines the maximum score possible and produces an alignment with that score.[1,4–8]

The scoring matrix is constructed from an input matrix, typically the mutation data matrix, and a matrix bias parameter, B, that is added to all terms of the input matrix. The net effect of adding B is that the score for any given alignment is increased by B times the number of positions where a residue is aligned with another residue. If B is increased, the alignment with the maximum score frequently changes to one with more residues aligned, and it therefore has a shorter overall length.

The maximum score, S, that can be achieved by an alignment of a pair of real sequences is compared with the distribution of maximum scores for a large number (usually 100) of random permutations of the two sequences. The mean and standard deviation of this approximately normal distribution are S_r and SD_r. The alignment score, A, is the number of standard deviations by which the maximum score for the real sequences exceeds the average maximum score for the random permutations: $A = (S - S_r)/SD_r$ (in SD units).

The alignment score is thus expressed in units of standard deviations from the mean of random scores. The statistic A will be normally distrib-

uted with a mean of zero and a standard deviation of 1.0 if the sequences being compared are unrelated, if the scores from the randomized sequences form a normal distribution, and if a sufficient number, N, of random scores is generated. The standard deviation of the mean is $1/\sqrt{N}$ and the standard deviation of the standard deviation is $1/\sqrt{2N}$. We have calculated alignment scores for randomized sequences and for real sequences that are unrelated. A chi-square test on the six intervals between -3 and $+3$ standard deviations from the mean indicates that, in this range, these scores reasonably ($\chi^2 = 1.7$) fit a normal distribution.[4]

The alignment score in SD units is not proportional to evolutionary distance, and its value is affected by the parameters of a given comparison. However, it is very useful for making statements regarding the probability that the similarity observed could have occurred by chance. The probability that a score as high as that from the real sequences could have been obtained in a comparison of randomized or unrelated sequences can be determined from the cumulative standardized normal distribution, as shown in Fig. 1. The results of a comparison of human α_1-microglobulin[32] and bovine lactoglobulin[39] are shown in Fig. 6. The ALIGN program is our most sensitive tool for comparing sequences of similar length with no long internal duplications or rearrangements. Because the amount of time and core memory needed for the algorithm are approximately proportional to the product of the lengths of the two sequences, it can become impractical to use this program for very long sequences.

There are several situations in which one is tempted to select portions of two sequences for comparison with, for instance, program ALIGN.

1. The longer sequence appears to consist of domains, one or more of which may be homologous with the shorter sequence, and one domain is selected for the comparison.
2. One sequence is much shorter and appears homologous with a part of a longer sequence, and, therefore, only some of the longer sequence is used.
3. A portion of one sequence appears to be homologous with a portion of the other sequence, and only a part of each sequence is selected for comparison.

As a general rule, one must remember that one has biased the results if the selection of the portions to be compared is based entirely on sequence similarity, especially if the similarity was discovered by examining a large number of sequences. For instance, if our database is searched with a segment 50 residues long, the test segment is compared with nearly

[39] G. Braunitzer, R. Chen, B. Schrank, and A. Stangl, *Hoppe-Seyler's Z. Physiol. Chem.* **354**, 867 (1973).

```
                MGHUA1 - Alpha-1-microglobulin - Human
                LGBO   - Lactoglobulin - Bovine and goat

          1         5         10        15        20        25        30
          '         '         '         '         '         '         '
MGHUA1    G P V P T P P D N I Q V Q E N F N I S R I Y G K W Y N L - -
LGBO      - L I V T - - - - - Q T M K G L D I Q K V A G T W Y S L A M
Common              T                   Q             I       G   W Y   L

         31        35        40        45        50        55        60
          '         '         '         '         '         '         '
MGHUA1    - - - - - K I M D R M T V S T L V L G E G - - - A T E A E I
LGBO      A A S D I S L L D A Q S A P L R V Y V E E L K P T P E G D L
Common                    D                   V     E             E

         61        65        70        75        80        85        90
          '         '         '         '         '         '         '
MGHUA1    S M T S T R W R K G V C E E T S G A Y E K T D T D G K F L Y
LGBO      E I L L Q K W E N G E C A Q K K I I A E K T K I P A V F K I
Common                W   G   C                   E K T           F

        91        95        100       105       110       115       120
         '         '         '         '         '         '         '
MGHUA1    H K S K W N I T M E S Y V V H T N Y D E Y A I F - L T K F S
LGBO      D A L N E N K V L - - - V L D T D Y K K Y L L F C M E N S A
Common              N             V     T   Y     Y     F

       121       125       130       135       140       145       150
         '         '         '         '         '         '         '
MGHUA1    R H T G P I T A K L Y G R A P Q L R E T L L Q D F R V V A Q
LGBO      E P E Q S L A C Q C L V R T P E V D D E A L E K F D K A L K
Common                          R   P             L     F

       151       155       160       165       170       175       180
         '         '         '         '         '         '         '
MGHUA1    G V G I P E D S I F T M A D R G E - C V P G E Q Q P E P I
LGBO      A L P M H I R L S F N P T Q L E E Q C - - - - - - - - H I
Common                      F             E   C                   I

          Mutation Data Matrix (250 PAMs) + 6
           32 Identities out of 150 possible matches between residues
            7 Breaks, penalty for a break = 6

          Total score (R) = 1064
          100 random runs
          Mean score (M) = 960.65
          R-M = 103.35
          Standard deviation = 17.72
          Alignment score = 5.83
```

FIG. 6. Alignment of human α_1-microglobulin[32] and bovine lactoglobulin.[39] These sequences are 77% different, and the alignment score is 5.8, near the limit of detection. A bias of 6 and a gap (break) penalty of 6 were used. For distantly related sequences, there are frequently several alignments that obtain the same best alignment score. The alignment score can be significantly high even when there are many equally good alignments.

180,000 segments. Even if these segments are all unrelated, the highest scoring of these will give a probability of about 10^{-5}, corresponding to a score of 4.3 SD, if compared with the test segment using ALIGN or RELATE.

It is preferable, particularly when trying to demonstrate a very distant relationship, to choose the sequence portions to be compared by indepen-

dent criteria such as domains that have been identified by X-ray crystallography,[19] duplicated regions within a sequence, exons that have been identified by nucleotide sequencing, or fragments shown to have a particular activity. When choosing segments to compare, there are many more possibilities than when choosing whole sequences. The probability of 10^{-6} used for superfamily clustering is not appropriate; a lower value will be required, depending on the method of selection employed.

Segment Comparison Scores: Program RELATE

A second computer method for detecting unusual similarity between sequences compares all possible segments of a given length from one sequence with all segments of the same length from the second sequence.[1-3] This method is particularly useful when it is not apparent which residues correspond; for example, if the sequences are of very different length, if the segments of one sequence are not colinear with related segments in the other, and if duplications are to be detected within a single sequence. Very long sequences can be compared, although the time required is proportional to the product of the lengths. A segment score is accumulated from the pair scores of the amino acids occupying corresponding positions within two segments. A scoring matrix is supplied by the user, as before. For example, if the length of the segments used is 25 and the total lengths of the two proteins compared are 100 and 120 amino acids, then 7296 scores will be tabulated. At most, 76 of these will come from comparisons of corresponding segments. As with program ALIGN, numerical properties of the distribution of scores are determined for the real sequences and for a number of randomized sequences. A RELATE score (in SD units) is calculated as the difference between the value determined for the real sequences and the average value determined from the many pairs of randomized sequences, divided by the standard deviation of the values from the randomized sequences. The probability of occurrence of a particular RELATE score by chance can be found in Fig. 1. We use two different kinds of numerical property. The first is the RELATE Magnitude, the average magnitude of a predetermined number of highest scores, usually chosen to be the number of scores to be expected if the sequences are related (i.e., 76 in the above example).

The program also calculates a second kind of numerical property, the RELATE Count score. A whole spectrum of Count scores, SC(N), in SD units, is determined for each computer run (Table V).[23,40] For each of these, the numerical property is the number of scores at or above the Nth scoring interval. One can examine the spectrum to find the maximum

[40] A. Kurosky, D. R. Barnett, T.-H. Lee, B. Touchstone, R. E. Hay, M. S. Arnott, B. H. Bowman, and W. M. Fitch, *Proc. Natl. Acad. Sci. U.S.A.* 77, 3388 (1980).

TABLE V
SPECTRUM OF COUNT SCORES FROM PROGRAM RELATE[a]

Cumulative values					
Lowest score of range	No. (real)	Mean No. (random)	Difference	SD (random)	Count score[b] (SD units)
0	2885	2970.58	−85.58	223.35	−0.383
10	544	532.19	11.81	64.23	0.184
20	175	67.93	107.07	20.33	5.267
30	86	5.59	80.41	5.34	15.050
40	52	0.29	51.71	1.30	39.626
50	14	0.03	13.97	0.22	62.730
60	5	0.00	5.00	0.00	—

[a] Human haptoglobin[40] and bovine trypsinogen[23] were compared.
[b] The number of matches with scores ≥20 clearly exceeds that expected by chance.

SC(N) and the corresponding number of segments with higher scores from the real sequences. For related sequences, sometimes there are only a few unexpectedly high-scoring segments, which are the result of preferential conservation of a small active site. Sometimes, as in the case of extensively conserved sequences, all of the corresponding segments get scores above the rest of the distribution. On occasion, the maximum value corresponds to a surplus of hundreds of scores above the expected score, but all well below the maximum score. This situation occurs when the sequences are very distantly related and when they both have evolved through many repetitions of similar segments. No single comparison may be outstanding, but the numerous intercomparisons of the repeated segments yield a bulge in the upper tail of the distribution.

Several other kinds of output are obtained from this program, including an ordered list of many segment comparisons giving the highest scores, from which the regions of the sequences showing unusual similarity can be identified. From this list, a plot of the segments that match best in the two sequences can be made by hand or by computer,[41–43] as shown in Fig. 7.[44,45] Longer contiguous sequences that match appear as diagonal lines on this figure and are easily interpreted by the human eye.

[41] A. J. Gibbs and G. A. McIntyre, *Eur. J. Biochem.* **16,** 1 (1970).
[42] J. V. Maizel, Jr., and R. P. Lenk, *Proc. Natl. Acad. Sci. U.S.A.* **78,** 7665 (1981).
[43] W. C. Barker, L. K. Ketcham, and M. O. Dayhoff, *in* "Atlas of Protein Sequence and Structure" (M. O. Dayhoff, ed.), Vol. 5, Suppl. 3, pp. 359–362. National Biomedical Research Foundation, Washington, D.C., 1979.
[44] T. Hase, H. Matsubara, and M. C. W. Evans, *J. Biochem.* (*Tokyo*) **81,** 1745 (1977).
[45] M. Tanaka, T. Nakashima, A. M. Benson, H. F. Mower, and K. T. Yasunobu, *Biochemistry* **5,** 1666 (1966).

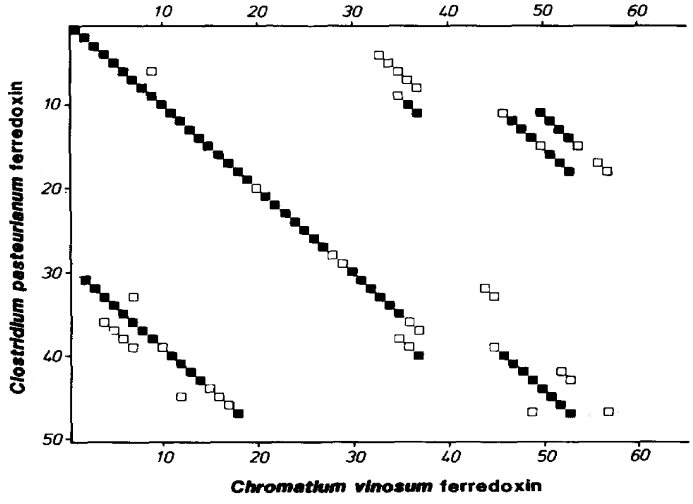


FIG. 7. The highest scoring segment pairs of *Chromatium vinosum*[44] and *Clostridium pasteurianum*[45] ferredoxins determined by program RELATE. All segments eight residues long were compared using the mutation data matrix. Scores ≥22 are shown by solid boxes and those from 18 to 21, by open boxes. It is clear from the main diagonal line that the sequences match from one end to the other, but with an insertion in the *Chromatium* sequence between residues 37 and 43. The short diagonal lines above and below the main diagonal reflect duplicated sequences in the molecules.

This program is particularly useful for the comparison of a sequence with itself (omitting the comparison of identical segments) to show the existence and location of single or multiple repeated segments. We have developed a screening procedure to detect major duplications, using the RELATE Magnitude property, and applied it to the families in the database.[43,46] Some of the most pronounced duplications are shown in Table VI. The number of highest scores to be used was determined for each protein from the segment length, s, and the total length of the sequence, L: $L/2 - s + 1$. This expression is equal to the number of scores to be expected from comparisons of corresponding segments if the sequence has exactly doubled. A protein that has duplicated will have many high scores for segments that are displaced by half of its total length. A protein with a prominent 10-residue periodicity will have many high scores at displacements of 10, 20, 30, etc. Such additional matches also appear on the plot, as shown in Fig. 7 for the ferredoxin molecule.

Although this procedure provides an easy and straightforward criterion for internal duplication, it may fail to detect duplications in certain

[46] W. C. Barker, L. K. Ketcham, and M. O. Dayhoff, *J. Mol. Evol.* **10**, 265 (1978).

TABLE VI
SELECTED PROTEINS WITH INTERNAL DUPLICATIONS

Protein (source)	RELATE score[a] (SD units)	Length of protein	Approximate length of repeat	Number of repetitions	Percent repetitions
Cytochrome c_7 (*Desulfuromonas acetoxidans*)	4.1	68	18	3	79
Cytochrome c_3 (*Desulfovibrio vulgaris*)	3.7	107	50[b]	2	93
Ferredoxin (*Clostridium pasteurianum*)	6.8	55	28	2	100
Rubredoxin (*Pseudomonas oleovorans*)	3.5	174	55	2	63
High-potential iron-sulfur protein (*Rhodopseudomonas gelatinosa*)	6.1	74	14	3	57
Prothrombin (bovine)	10.8[c]	582	79	2	27
Plasminogen (human)	44.5	790	79	5	50
Ovomucoid (Japanese quail)	17.4	186	59	3	95
Protease inhibitor, submandibular gland (dog)	7.4	115	54	2	94
Protease inhibitor, Bowman–Birk (soybean)	3.9	71	28	2	79
Calcium-dependent regulator protein (bovine)	13.0	148	74[b]	2	100
Tropomyosin α chain (rabbit)	5.8	284	42[b]	7	100
Immunoglobulin μ chain C region (human)	19.1	452	108	4	96
Serum albumin (human)	12.1[d]	584	195[b]	3	100
Histone H3 (bovine)	3.6	135	9 13	3 2	39

[a] All scores were determined using the mutation data matrix, a segment length of 15 or 20, and 100 random runs.
[b] This major repeating unit shows evidence of repetitious structure within itself.
[c] Score based on testing residues 1–323.
[d] Score based on testing residues 1–500.

cases. If a duplication involves only a small fraction of the sequence, it may not be detected unless it is composed of amino acids that are usually highly conserved in proteins. We have purposely designed the procedure to detect major duplication and extensive periodicity. If too many changes have occurred in the sequence, an ancestral duplication may not be detected. For detecting ancient duplications, we use the mutation data matrix and segments of at least 20 residues. When looking for small, recent duplications, we use the unitary matrix and a length of 5 or 10 residues.

Scoring Matrices

The methods for detecting distant relationships described above depend on an amino acid pair score matrix.[22] The simplest of these, the unitary matrix (UM), assigns a value of +1 to identical residues and 0 to nonidentical ones. A slightly more complicated scoring system reflects the maximum number of identities in the nucleotides of the genes coding for the proteins. Identical amino acids obtain a score of 3; those for which two nucleotides could be identical, 2; one nucleotide, 1; and 0 if no nucleotides are ever shared in the codons for the amino acids. We refer to this as the genetic code matrix (GCM). In 1971, a matrix based on alternative amino acids (AAAM) at each position in alignments of groups of related sequences was derived by McLachlan.[47] In 1968, we described a method to derive scoring matrices for sequences at any evolutionary distance, based on amino acid replacements between present-day sequences and those inferred as common ancestors on evolutionary trees.[48] The residues that did not change and the relative exposure of the sequences to mutational change were taken into account. In 1978 we rederived the mutation data matrix (MDM) on the basis of 1572 mutations observed in families of closely related sequences.[49,50]

These raw data were converted to a Mutation Probability Matrix. An element of this matrix gives the probability that the amino acid in column j will be replaced by the amino acid in row i after a given evolutionary interval, in this case 1 percent accepted mutation (PAM) in a sequence of

[47] A. D. McLachlan, *J. Mol. Biol.* **61**, 409 (1971).

[48] M. O. Dayhoff and R. V. Eck, "Atlas of Protein Sequence and Structure 1967–1968," pp. 33–41. National Biomedical Research Foundation, Silver Spring, Maryland, 1968.

[49] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt, *in* "Atlas of Protein Sequence and Structure" (M. O. Dayhoff, ed.), Vol. 5, Suppl. 3, pp. 345–352. National Biomedical Research Foundation, Washington, D.C., 1979.

[50] R. M. Schwartz and M. O. Dayhoff, *in* "Origin of Life: Proceedings of the Second ISSOL Meeting, Fifth ICOL Meeting" (H. Noda, ed.), pp. 457–469. Center for Academic Publications Japan/Japan Scientific, Tokyo, 1978.

average amino acid composition. There is a different mutation probability matrix for each evolutionary distance. The 1-PAM matrix can be multiplied by itself N times to yield a matrix that predicts the amino acid replacements to be found after N PAMs of evolutionary change in a sequence of average composition. To derive a scoring matrix from a probability matrix, each element, representing the probability of an exchange due to a mutation, is divided by the probability that two amino acids will be found by chance. The log of this ratio gives the element of the mutation data scoring matrix (MDM). Using programs ALIGN and RELATE and distantly related sequences, we compared a number of matrices including UM, GCM, AAAM, and mutation data matrices of different evolutionary distances. From these tests we concluded that all these matrices give very high scores with closely related sequences but that the MDM matrix for 250 PAMs (Fig. 3) is the best matrix for detecting distantly related sequences.[22]

Comments

The methods described above or procedures very similar to them have been used by ourselves and others to elucidate a number of surprising relationships, among which are the catalytic chain of bovine cyclic AMP-dependent protein kinase and the *src* gene products of Rous avian and Moloney murine sarcoma viruses[51]; α_1-microglobulin, α_{2u}-globulin, lactoglobulin, and plasma retinol-binding protein[52,53]; antithrombin-III, α_1-antitrypsin, and ovalbumin[54]; epidermal growth factor and the light chain of coagulation factor X[55]; the leech protease inhibitor eglin C and potato chymotrypsin inhibitor I, A chain[17]; and apolipoproteins A-I, A-II, C-I, and C-III.[56] The reader can, with these procedures, readily verify the above relationships or possibly discover new ones.

Acknowledgments

This work was partially supported by NIH Grant GM-08710 and NASA Contract NASW3317.

[51] W. C. Barker and M. O. Dayhoff, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 2836 (1982).
[52] R. D. Unterman, K. R. Lynch, H. L. Nakhasi, K. P. Dolan, J. W. Hamilton, D. V. Cohn, and P. Feigelson, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 3478 (1981).
[53] P. Feigelson, personal communication, 1981.
[54] L. T. Hunt and M. O. Dayhoff, *Biochem. Biophys. Res. Commun.* **95**, 864 (1980).
[55] C. L. Young, W. C. Barker, C. M. Tomaselli, and M. O. Dayhoff, *in* "Atlas of Protein Sequence and Structure" (M. O. Dayhoff, ed.), Vol. 5, Suppl. 3, pp. 73–93. National Biomedical Research Foundation, Washington, D.C., 1979.
[56] W. C. Barker and M. O. Dayhoff, *Comp. Biochem. Physiol.* **57B**, 309 (1977).